Name: Graham Northrup
Section: 1

TTIC 31050 # 4

Introduction

The goal of this project was to assess the Q3 scores of two secondary structure tools, SPIDER2 and RAPTORX. Each uses a different approach to 'guess' the secondary structure of the given amino acid sequence. the DSSP uses 8 different markers for secondary structure, however the Q3 score and most algorithms are interested in the 3 states sheet, helix, and coil. Our 8 to 3 condensation is discussed right below in the implementation section.

The code written this week was mostly file processing and I/O operations. This is because SPIDER2, RAPTORX, and DSSP all have online tools to access which means I can simply acquire the files and be 100% the implementation is correct.

Implementation

There aren't too many implementation notes for this assignment. A large one is our 8 to 3 algorithm which decides that DSSP codes 'E' and 'B' go to sheet('E'), 'G' and 'H' go to helix('H'), and everything else goes to coil ('C'). There are other algorithms out there but Cuff and Barton 1999 showed that these choices have minimal effects on the score.

An important note here is that I decided to assign blanks (' ') to coil ('C') as well. This is perhaps a very simplifying assumption, but it was seen some other places while doing some reading for the project so I decided to use it as it makes the calculations much more straightforward.

Another assumption worth noting is that when the lengths of a ground truth and guessed structure were of different lengths for one reason or another, my code simply aligns them at the beginning and scores until one of the strings runs out of letters.

Results

Table 1: The Q3 scores for each algorithm on sequences 1-8

	5a7dB	5ereA	5j5vA	5ko9A	5jmuA	5fjlA	5j4aA	5aotA
SPIDER2	0.69	0.75	0.72	0.57	0.42	0.43	0.30	0.32
RAPTORX	0.71	0.84	0.77	0.52	0.37	0.44	0.36	0.39

We can see the algorithms perform rather well on most sequences, with RAPTORX having a high score of over 80%. We do also see scores in the 30% range as well however, meaning these algorithms perform better on certain sequences than others. The SPIDER2

Table 2: The Q3 scores for each algorithm on each sequences 9-15

	4ympA	5a7dL	5jmbA	5kkpA	5j4aB	5j5vC	5j5vB
SPIDER2	0.46	0.66	0.61	0.37	0.54	0.74	0.62
RAPTORX	0.44	0.71	0.60	0.38	0.66	0.72	0.60

score performed with an average score of 0.55, where RAPTORX's average Q3 score was 0.57. We can also guess the algorithms to have similar averages by comparing each algorithm sequence-wise. While the each algorithm performed better and worse on certain sequences, these algorithms were largely the same for the same sequence (perhaps implying certain proteins lend their secondary structures more easily than others).

These scores however are much lower than the scores we see in table 1 of Yang et. al 2016. For each algorithm the Q3 score hovers around 80%. I am inclined to attribute this to an implementation error, as perhaps making a more concerted effort to align the structures of different lengths would produce a better score, as well as a more nuanced assumption for how to fill the blanks in the DSSP ground truth.

While a potentially large difference, the closeness of the algorithms to each other in terms of score could perhaps mean there is another underlying artifact of the scoring or algorithm that is depressing both scores by approximately the same amount.

Conclusion

The two algorithms tested performed similarly across the 15 sequences used. We suspect some underlying artifact is responsible for the difference in the scores calculated here and the scores seen in the literature; however it is uncertain at this time. A next effort could be to perhaps entertain different implementation choices, as well as incorporate other secondary structure tools to see how those compare to the ones we have already used here.